Data integrative Bayesian inference for mixtures of regression models

Aflakparast, Mehran; de Gunst, M.C.M.

published in
Journal of the Royal Statistical Society: Series C (Applied Statistics)
2019

DOI (link to publisher)
10.1111/rssc.12346

document version
Publisher's PDF, also known as Version of record

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2019)
68, *Part* 4, *pp.* 941–962

Data integrative Bayesian inference for mixtures of regression models

Mehran Aflakparast and Mathisca de Gunst

Vrije Universiteit Amsterdam, The Netherlands

[Received April 2018. Final revision February 2019]

Summary. Modern data collection techniques, which often produce different types of relevant information, call for new statistical learning methods that are adapted to cope with data integration. In the paper Bayesian inference is considered for mixtures of regression models with an unknown number of components, that facilitates data integration and variable selection for high dimensional data. In the approach presented, named data integrative mixture of regressions, data integration is accomplished by introducing a new data allocation scheme that summarizes additional data in the form of an informative prior on latent variables. To cope with high dimensionality, a shrinkage-type prior is assumed on the regression parameters, and *a posteriori* variable selection is conducted based on Bayesian credible intervals. Posterior estimation is achieved via a Markov chain Monte Carlo algorithm. The method is validated through simulation studies and illustrated by its performance on real data.

Keywords: Bayesian lasso; Data integration; Markov chain Monte Carlo algorithm; Mixture regression

1. Introduction

High dimensional data provided by the rapid progress in data collection techniques have motivated scientists of various fields to develop more computationally powerful methods for data analysis. The complexity of data structures and the need to combine multiple relevant sources of data introduce new challenges in contemporary statistical learning. Mixture models are among the most mature statistical methods that have been of strong and sustained interest in applications where the complexity of the system of interest is due to heterogeneity of the population; see for example Goldfeld and Quandt (1976), Everitt and Hand (1981) and McLachlan and Peel (2000) for comprehensive discussions. An important application of mixture models concerns relating an independent random variable with a finite mixture distribution to a set of covariates. This provides a generalization of mixture models which is known as *finite mixture of regression* (FMR) models that was first introduced by Quandt and Ramsey (1978); for a review see Wedel and Kamakura (2012) and Hurn *et al*. (2003). In this study we develop a flexible Bayesian method for fitting FMR models. This method incorporates prior knowledge on sample clustering from additional sources of data in a novel way, which makes it superior to existing methods. Additionally, we address long-standing difficulties in fitting FMR models such as estimation of the number of mixture components, high dimensionality problems and clustering inaccuracy.

In regression modelling for high dimensional data, redundancy of covariates is generally addressed by regularization and variable-selection strategies (Tibshirani, 2011; Williams, 1995;

Address for correspondence: Mathisca de Gunst, Department of Mathematics, Vrije Universiteit Amsterdam, 108 HV Amsterdam, The Netherlands.
E-mail: degunst@cs.vu.nl

 0035–9254/19/68941

Xie *et al.*, 2018). Accordingly, in the context of FMR models, it is crucial to retain only the most significant covariates in each subpopulation to avoid overfitting and to strengthen model interpretability. Despite substantial literature on variable selection in mixture models (see for example Tadesse *et al.* (2005), Maugis and Martin-Magniette (2009) and Yau and Holmes (2011)) to date the literature on variable selection for FMR models is still limited. Classical information-theory-based approaches such as those using the Akaike information criterion AIC or the Bayesian information criterion BIC, although being straightforward, are computationally expensive (Khalili, 2011). Most of the research regarding variable selection in FMR models has appeared in the context of *mixture-of-experts* models in which the mixing probabilities are assumed to be a function of covariates (Jacobs *et al.*, 1991; Jordan and Jacobs, 1994). General discussions and examples of different approaches that are related to variable selection in mixture-of-experts models can be found in Jacobs *et al.* (1997), Gupta and Ibrahim (2007), Villani *et al.* (2009), Chung and Dunson (2009) and Tran *et al.* (2012). In this study, we avoid the additional cost of estimating unknown parameters corresponding to mixing probabilities and confine our approach to merely models that are dependent on predictors only through the component means.

We take a Bayesian approach and address the high dimensionality problem by means of a componentwise lasso-type shrinkage probability on the regression parameters. This makes our method comparable with those introduced in Khalili and Chen (2007) and Städler *et al.* (2010), which employ the same type of penalty but in a frequentist context. Khalili and Chen (2007) suggested numerical solutions for maximization of the l_1-penalized likelihood function by replacing the penalty with a local quadratic approximation. In a similar frequentist fashion, Städler *et al.* (2010) presented a different parameterization of the non-convex log-likelihood function for FMR models combined with a generalized block co-ordinate descent expectation–maximization algorithm under the name of FMRLasso. These methods alleviate the computational burden to a significant degree and enjoy favourable statistical properties. However, data integration, which is highly demanded in modern applications, is not easily feasible in these types of methodologies. Moreover, the uncertainty about the values of the tuning parameters and the number of model components can lead to inaccurate estimation of parameters. Since these quantities are commonly determined based on fitting models over a grid of predefined values and comparing goodness-of-fit quantities, this may bring additional computational problems.

Data clustering and clusterwise parameter estimation are the two major parts of mixture modelling. To improve performance on both sides, one can consider making use of available additional information such as cluster information or similarity measurements from other sources of data. This requires methods with enough flexibility to let the data themselves determine the number of mixture components and to take into account the additional information.

With a Bayesian approach additional information can conveniently be incorporated into prior distributions. Richardson and Green (1997) introduced a reversible jump Markov chain Monte Carlo (MCMC) algorithm for estimating the mixture model parameters as well as the number of components in a Bayesian framework. Stephens (2000) proposed an alternative Bayesian MCMC approach based on marked point processes. However, these approaches have limited flexibility in how to incorporate additional information and, therefore, may make use of only a small part of the available information.

Alternatively, for clustering, non-parametric Bayesian approaches can be used. Dirichlet process mixture models are popular random-partitioning models that allow the number of mixture components to grow by the data (Antoniak, 1974). A prevalent way of representing Dirichlet process mixtures is through a so-called Chinese restaurant process (CRP), in which a new data point is allocated to one of the existing components with a conditional probability that depends on the size of the component, or to a new component with some fixed probability. In

Neal (2000) and Rasmussen (2000) CRP-based Bayesian clustering methods are considered. A useful feature of these methods is that for estimating the unknown number of components it is not necessary *a priori* to limit the number of components of the mixture model to be finite. However, for many applications, like clustering of genetic data, the principle of 'the rich gets richer' that underlies this type of methods is not appropriate. Next to this, the data exchangeability assumption makes CRP-based strategies inadequate for situations where clustering naturally depends on certain characteristics of the data points. To deal with either of these issues several alternative random-partitioning methods have been proposed (see, for example, Rasmussen and Ghahramani (2002), Müller and Quintana (2010) and Blei and Frazier (2011)). In this paper we simultaneously address both issues.

We introduce a Bayesian clustering method in the spirit of Neal (2000) and Rasmussen (2000) with a new CRP-based data allocation strategy that simultaneously deals with both issues and also facilitates the integration of additional clustering or similarity information. Our Bayesian estimation procedure is implemented via an MCMC algorithm with a Gibbs sampler at the heart of it.

The remainder of this paper is organized as follows. In Section 2 the model and the estimation problem are introduced. In Section 3 the prior distributions for the componentwise regression parameters are given and the new data allocation scheme is presented, whereas Section 4 concerns the posterior distributions. In this section we also present our hybrid MCMC algorithm, named data integrative mixture of regressions (DIMR), and discuss variable selection based on credible intervals. In Section 5 we evaluate our method with respect to different aspects such as practical consistency, convergence of regression parameters, clustering accuracy and comparison with FMRLasso on simulated data. Then in Section 6 we illustrate our method by its application to a real stomach cancer data set. We conclude in Section 7 with a discussion on limitations and possible extensions of our work.

2. Model and estimation problem

We consider, for $i = 1, \ldots, n$, a linear regression model with univariate, normally distributed response variable Y_i and corresponding p-variate explanatory vectors $X_i \in \mathcal{R}^p$. We always work conditionally on $X_i = x_i$. We assume that the data comprise K unknown components (or clusters). It is further assumed that given the clustering the Y_i are independent, and that the way in which the covariates contribute to the response variable is the same within a cluster but differs between clusters. For each Y_i the probability that it belongs to the kth cluster is π_k, $k = 1, \ldots, K$, with $\Sigma_{k=1}^{K} \pi_k = 1$. The following finite mixture of (linear) regressions FMR model is considered:

$$Y_i | x_i, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\pi} \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i^{\mathrm{T}} \beta_k, \sigma_k^2), \qquad i = 1, \ldots, n, \tag{1}$$

where

$$\begin{aligned} \boldsymbol{\beta} &= (\beta_1, \ldots, \beta_K)^{\mathrm{T}}, \\ \beta_k &= (\beta_{k1}, \ldots, \beta_{kp})^{\mathrm{T}}, \end{aligned} \tag{2}$$

$$\begin{aligned} \boldsymbol{\sigma}^2 &= (\sigma_1^2, \ldots, \sigma_K^2)^{\mathrm{T}}, \\ \boldsymbol{\pi} &= (\pi_1, \ldots, \pi_K)^{\mathrm{T}}, \end{aligned} \tag{3}$$

and $\mathcal{N}(x_i^{\mathrm{T}} \beta_k, \sigma_k^2)$ stands for the univariate normal distribution with mean $x_i^{\mathrm{T}} \beta_k$ and variance σ_k^2. This means that β_k denotes the p-vector of regression coefficients corresponding to component k.

We use capital letters to denote random variables or random vectors, and small letters for their realizations, bold type for quantities belonging to the full model and subscript or superscript k for quantities belonging to the kth cluster. In what follows we shall use $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ for the vector of all response variables, $\mathbf{X} = (X_1, \ldots, X_n)^{\mathrm{T}}$ for the $n \times p$ matrix of explanatory variables and n_k for the size of component k, i.e. for the number of Y_i that belong to the kth component. Furthermore, $Y^k = (Y^k_1, \ldots, Y^k_{n_k})^{\mathrm{T}}$ will denote the n_k-vector of those Y_i that belong to component k, and $X^k = (X^k_1, \ldots, X^k_{n_k})^{\mathrm{T}}$ its corresponding $n_k \times p$ matrix of explanatory variables.

Our aim is to estimate the unknown values of the parameters $\boldsymbol{\beta}$, $\boldsymbol{\sigma}^2$ and $\boldsymbol{\pi}$, and the number of components K. For this we take a Bayesian approach and approximate the posterior distributions by means of MCMC sampling.

For parameter estimation in mixture models one commonly makes use of a missing data approach, which we also adopt in this study. This not only simplifies computations but also facilitates estimation of the component memberships of the Y_i. We thus introduce a vector of n independent missing, or latent, variables $\mathbf{Z} = (Z_1, \ldots, Z_n)^{\mathrm{T}}$, where Z_i represents the component membership of data point i, so that

$$p(Z_i = k) = \pi_k, \qquad k = 1, \ldots, K.$$

Because, conditionally on $\mathbf{Z} = \mathbf{z}$, $Y_1, \ldots, Y_n$ are independent, the n_k are known and

$$Y_i | x_i, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2 \sim \mathcal{N}(x_i^{\mathrm{T}} \beta_{z_i}, \sigma^2_{z_i}), \qquad i = 1, \ldots, n, \tag{4}$$

we have that conditionally on $\mathbf{Z} = \mathbf{z}$ the data likelihood can written as

$$p(\mathbf{y} | \mathbf{x}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2) = \prod_{k=1}^{K} p(y^k | x^k, \mathbf{z}, \beta_k, \sigma^2_k), \tag{5}$$

where

$$p(y^k | x^k, \mathbf{z}, \beta_k, \sigma^2_k) = (2\pi\sigma^2_k)^{-n_k/2} \exp\left\{ -\frac{1}{2\sigma^2_k} (y^k - x^k \beta_k)^{\mathrm{T}} (y^k - x^k \beta_k) \right\}, \tag{6}$$

with y^k and x^k denoting the realizations of the response vector Y^k and corresponding covariate matrix X^k respectively of the data points belonging to component k.

Furthermore, given the set of parameters $\boldsymbol{\beta}$, $\boldsymbol{\sigma}^2$ and $\boldsymbol{\pi}$, the complete-data likelihood can be written as

$$p(\mathbf{y}, \mathbf{z} | \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\pi}) = p(\mathbf{y} | \mathbf{x}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2)\, p(\mathbf{z} | \boldsymbol{\pi}). \tag{7}$$

We remark that, in the model formulation above, the unknown number of components K is taken to be fixed and finite, whereas in the data allocation part of our estimation procedure the number of components is, similarly to the approach that was considered in Neal (2000) and Rasmussen (2000), in principle not limited to be finite. The finite number K that we aim to estimate should be interpreted as the number of components containing the n data points, which naturally is a finite number.

Finally, we introduce the notation for the additional similarity information. It is assumed that this additional information is available in the form of a symmetric $n \times n$ matrix $\mathbf{S} = (S_{ii'})$, where $S_{ii'}$ is a non-negative random variable representing the additional similarity information between data points i and i'. The matrix $\mathbf{S}$ will be referred to as the *additional data* or as the *similarity matrix*. We emphasize that $S_{ii'}$ should not have been computed from Y_i and $Y_{i'}$, but that $\mathbf{S}$ should originate from an additional source of data, such that $\mathbf{Y}$ and $\mathbf{S}$ can be assumed

to be independent. The measured value $\mathbf{s}$ of $\mathbf{S}$ will be incorporated in the prior distributions for the component memberships.

3. Prior distributions

In this section we define the prior distributions for the model parameters. We start with the regression parameters $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$. Next, we introduce a new clustering scheme that defines the prior distribution for the (conditional) component memberships.

3.1. Priors for regression parameters

From expression (4), we see that, given the component membership vector $\mathbf{Z}$, the estimation problem turns into fitting K independent regression models. As our focus is on high dimensional problems, for each k the regression coefficient vector β_k is assumed to be sparse, in the sense that within each component only a few covariates contribute to the variability of the response variable. We use shrinkage estimation of the regression coefficients and apply the Bayesian lasso procedure of Park and Casella (2008). Accordingly, to design an appropriate Gibbs sampler, we define, for $k = 1, \ldots, K$, a hyperparameter $\tau_k^2 = (\tau_{k1}^2, \ldots, \tau_{kp}^2)$ with independent and identically distributed elements and assume that

$$\begin{aligned} \beta_k|\sigma_k^2, \tau_k^2 &\sim \mathcal{N}_p(\mathbf{0}_p, \sigma_k^2 D_k), & D_k &= \operatorname{diag}(\tau_{k1}^2, \ldots, \tau_{kp}^2), \\ \tau_{kj}^2 &\sim \text{exponential}(\lambda_k^2/2), & j &= 1, \ldots, p, \end{aligned} \tag{8}$$

with $\mathbf{0}_p$ being the p-dimensional vector with only 0s. Hence, the prior density of τ_k^2 is given by

$$p(\tau_k^2) = \prod_{j=1}^{p} \frac{\lambda_k^2}{2} \exp\left(- \frac{\lambda_k^2 \tau_{kj}^2}{2} \right). \tag{9}$$

It can be seen that, for each k and j and conditionally on σ_k^2, we have independent double-exponential conditional prior distributions with location parameter 0 and scale parameter $\sqrt{(\sigma_k^2/\lambda_k^2)}$ for the components β_{kj}. This yields

$$p(\beta_k|\sigma_k^2) = \prod_{j=1}^{p} \frac{\sqrt{\lambda_k^2}}{2\sqrt{\sigma_k^2}} \exp\left(- \frac{\sqrt{\lambda_k^2}|\beta_{kj}|}{\sqrt{\sigma_k^2}} \right). \tag{10}$$

Furthermore, we assume an inverse gamma distribution $p(\sigma^2)$ *a priori* for σ_k^2,

$$\sigma_k^2 \sim \mathcal{IG}(\omega, \eta), \tag{11}$$

with ω and η the shape and scale parameter respectively, and a gamma prior for the tuning parameter λ_k^2,

$$\lambda_k^2 \sim \mathcal{G}(r, \delta), \tag{12}$$

with r being the shape and δ the rate parameter. We fix the shape parameter ω and set it to 1 to avoid extremely small posterior variances for σ_k^2, and for η we assume an exponential prior with mean ϕ_y that is equal to the sample variance of the response variable. Finally, we set the hyperparameter δ equal to a value that is sufficiently larger than 0 to avoid computational problems (Park and Casella, 2008).

With the above set-up we have (conditional) conjugacy for β_k. Moreover, the fact that σ_k^2 is included in the prior distribution (10) prevents multimodality of the joint posterior distribution

of σ_k^2 and β_k. We note that the tuning parameters λ_k are assumed to be component specific, so that a common tuning parameter is used for regularization of all regression coefficients within a component. However, with a trade for more computational costs, one could also assume specific tuning parameters for each coefficient (for example see the adaptive lasso of Zou (2006)).

3.2. Prior for component membership probabilities

Here we propose our new data allocation strategy, the data integrative Chinese restaurant process (DICRP), which can be used as an alternative to the CRP of Antoniak (1974) in settings where it is not realistic to assume that the data points are exchangeable. This strategy also allows for the integration of additional, external, data sets that contain some kind of similarity information about the data points Y_i, thereby increasing the accuracy of the data clustering, and hence of the parameter inference.

Recall that $\mathbf{S} = (S_{ii'})$ denotes the similarity matrix representing the additional similarity information with $S_{ii'}$ being the non-negative similarity value between data points i and i', and that $\mathbf{s} = (s_{ii'})$ denotes its observed value. We assume, for $i = 1, \ldots, n$ and $k = 1, \ldots$, the following conditional distribution on the component memberships:

$$p(Z_i = k | z_{-i}, \mathbf{s}, \alpha) = \begin{cases} n^*_{-i,k} h_i(k)/c, & \text{if } k \text{ is an existing component,} \\ \alpha/c, & \text{if } k \text{ is a new component.} \end{cases} \tag{13}$$

Here z_{-i} is the $(n-1)$-vector obtained from $\mathbf{z}$ by deleting z_i, $\alpha > 0$, c is a normalizing constant and

$$n^*_{-i,k} = \sum_{i':i' \neq i} I_{\{s_{ii'} \geqslant T_i\}} I_{\{z_{i'} = k\}}. \tag{14}$$

The function $h_i(k)$ in expression (13) is of the form

$$h_i(k) = 1 + \sum_{i':i' \neq i} s_{ii'} I_{\{z_{i'} = k\}} \tag{15}$$

and represents the overall similarity of data point i with all other data points in component k. In the above equations, I denotes the indicator function, and T_i is a threshold value which depends on data point i. In the applications below we shall choose T_i to be the third quantile of the similarity values between the data point i and the rest of the data points. This choice ensures that a data point is more likely to end up in a component where it has high similarity with the majority of the other data points. However, other choices are possible also. Numerous similarity (or distance) measures for either of categorical or continuous attributes have been proposed and studied in different disciplines (see for example Boriah *et al.* (2008) and Cha (2007)). We also note that other forms of h_i can be used, but for our purpose the simple form (15) suffices.

Prior (13) is an extension of the conditional probability that is reached by taking limits as the number of components tends to ∞ in a finite mixture model. The derivation of similar types of prior distributions on the component memberships has been well explained in Neal (2000) and Rasmussen (2000).

Obviously, the number of mixture components is largely controlled by the choice of α, in that larger values lead to more components. To estimate the precision parameter α, we follow Escobar and West (1995) and assume a gamma prior with mean a and rate parameter b:

$$\alpha \sim \mathcal{G}(a, b). \tag{16}$$

Similarity between two data points is usually described as their distance. If this distance is small, there will be a high degree of similarity; if the distance is large, there will be a low degree of

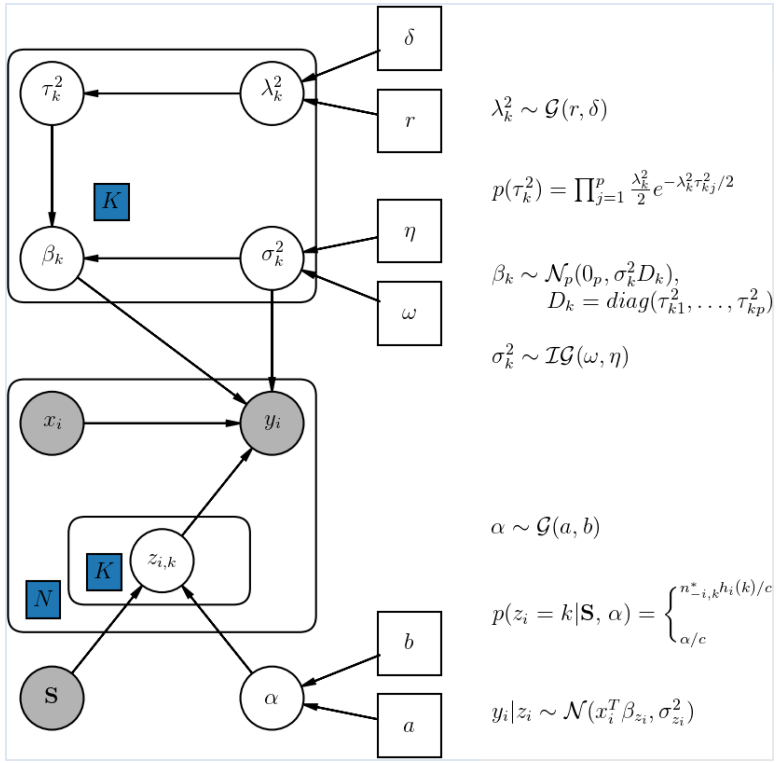

Fig. 1. Probabilistic graphical model representing DIMR with **S** the similarity matrix that is assumed available from additional sources of data

similarity. In our setting the judgement of similarity is based on magnitude and not orientation, so s_{ij} can take any non-negative value. This permits the DICRP to be equivalent to the original CRP in the case of having no additional data, where the similarity values are assumed to be 0 for all data points.

We have now presented all parts of our DIMR approach Fig. 1 shows its graphical model representation.

4. Full posterior conditionals

Given the component memberships $\mathbf{z} = (z_1, \ldots, z_n)^{\mathrm{T}}$, the full conditionals can be derived from the component likelihood (6) and the prior distributions of Section 3. Note that the priors in Section 3 are chosen to be conjugate so that their posteriors yield standard statistical distributions: the full conditionals for the regression coefficients are normally distributed, and the full conditional for the variance parameters are inverse gamma. We find for the regression parameters of component k:

$$\beta_k | x^k, y^k, \sigma_k^2, \tau_k^2 \sim \mathcal{N}_p\{A_k^{-1}(x^k)^{\mathrm{T}} y^k, \sigma_k^2 A_k^{-1}\}, \tag{17}$$

with $A_k = (x^k)^{\mathrm{T}} x^k + D_k^{-1}$, and for the variance parameter of component k we have

$$\sigma_k^2 | x^k, y^k, \beta_k, \tau_k^2 \sim \mathcal{IG}\left[\frac{n_k}{2} + \frac{p}{2} + \omega, \eta + \frac{1}{2}\{(y^k - x^k\beta_k)^{\mathrm{T}}(y^k - x^k\beta_k) + \beta_k^{\mathrm{T}} D_k^{-1} \beta_k\}\right]. \tag{18}$$

Conditional independence of $\tau_{k1}^2, \ldots, \tau_{kp}^2$ allows block updating from the following inverse normal distribution for the inverse parameters:

$$(\tau_{kj}^2)^{-1} | \beta_k, \sigma_k^2, \lambda_k \sim \mathcal{IN}(\mu_k', \lambda_k') \tag{19}$$

with $\mu_k' = \sqrt{(\lambda_k^2 \sigma_k^2 / \beta_{kj}^2)}$ the location and $\lambda' = \lambda_k^2$ the scale parameter.

Furthermore, we have

$$\eta | \sigma_k^2 \sim \mathcal{G}\left\{ K+1, \left(\frac{1}{\phi_y} + \sum_{k=1}^{K} \frac{1}{\sigma_k^2} \right)^{-1} \right\}, \tag{20}$$

and

$$\lambda_k^2 | \tau_k^2 \sim \mathcal{G}\left\{ p+r, \left(\delta + \sum_{j=1}^{p} \frac{\tau_{kj}^2}{2} \right)^{-1} \right\}. \tag{21}$$

The full conditional distribution for the hyperparameter α can be derived given the number of components K (which is implied by the fact that $\mathbf{z}$ is given), following the hierarchy that was introduced by Antoniak (1974):

$$\begin{gathered} \alpha | \zeta, K \sim \rho_\zeta \, \mathcal{G}\{a+K, b-\log(\zeta)\} + (1-\rho_\zeta) \, \mathcal{G}\{a+K-1, b-\log(\zeta)\}, \\ \zeta | \alpha \sim \text{beta}(\alpha+1, n). \end{gathered} \tag{22}$$

where

$$\frac{\rho_\zeta}{1-\rho_\zeta} = \frac{a+K-1}{n\{b-\log(\zeta)\}}.$$

Using expressions (5) and (13) we find that the conditional distribution of the latent variables satisfies

$$\begin{aligned} &p(Z_i = k \,|\, x_i, \, y_i, z_{-i}, \mathbf{s}, \alpha, \beta_k, \sigma_k^2) \\ &\qquad \propto \begin{cases} n_{-i,k}^* h_i(k) \, p(y_i | x_i, \beta_k, \sigma_k^2) & \text{if } k \text{ is an existing component,} \\ \int\int \alpha \, p(y_i | x_i, \beta_k, \sigma_k^2) \, p(\beta_k, \sigma_k^2) \mathrm{d}\beta_k \mathrm{d}\sigma_k^2 & \text{if } k \text{ is a new component,} \end{cases} \end{aligned} \tag{23}$$

where $p(y_i | x_i, \beta_k, \sigma_k^2)$ denotes the conditional density of Y_i which is specified by equation (4).

The double integral in expression (23) is not analytically tractable. We apply a Monte Carlo method as suggested by Neal (2000) and create a new component k', with parameter values $\beta_{k'}$ and $\sigma_{k'}^2$ generated from their priors and we replace the double integral by $p(y_i | x_i, \beta_{k'}, \sigma_{k'}^2)$. This enables Gibbs sampling from the posterior (23) in the case that a new component must be created. We note that a small component variance σ_k^2 can mask the contribution of the prior probabilities $n_{-i,k}^* h_i(k)$, which allows observations to be assigned to components other than k. However, when variances of components are more or less similar, the prior distribution can have a significant effect on the clustering.

We propose an MCMC algorithm, named DIMR, that consists of a Metropolis–Hastings sampler combined with a partial Gibbs sampler to update the component memberships according to our data allocation procedure DICRP described in Section 3.2, and together with the Bayesian lasso Gibbs sampler of Park and Casella (2008) to update the regression parameters. The algorithm iterates through the steps that are presented in Table 1. Note that in step 1 of Table 1 we use simple birth-and-death type updates for the number of components based on proposals from the prior densities of the component parameters. We have investigated other types of proposal, including a random walk, but the effect on the convergence or model perfor-

Table 1. DIMR algorithm

Step 1 (update number of components): for $i=1,\ldots,n$,
step 1.1, if data point i belongs to component k that has more than one occupant,
(a) create a new component k' and generate component parameters from the prior distributions,
(i) generate $\lambda^2_{k'} \sim \mathcal{G}(r, 1/\delta)$,
(ii) generate $\eta_{k'} \sim \exp(1/\phi_y)$,
(iii) generate $\sigma^2_{k'} \sim \mathcal{IG}(\omega, \eta_{k'})$,
(iv) generate $\tau^2_{k'j} \sim \exp(\lambda^2_{k'}/2)$, for $j=1,\ldots,p$,
(v) generate $\beta_{k'j} \sim \mathcal{N}(0, \sigma^2_{k'}\tau^2_{k'j})$, for $j=1,\ldots,p$
(b) retain the newly created component and put $z_i = k'$ with probability

$$\min\left\{1, \frac{\alpha}{n-1}\frac{p(y_i|x_i,\beta_{k'},\sigma^2_{k'})}{p(y_i|x_i,\beta_k,\sigma^2_k)}\right\};$$

step 1.2, if data point i belongs to component k that contains only one data point,
(a) propose a candidate component k' among the existing components with probability $n^*_{-ik'}/\Sigma_{k=1}^K n^*_{-ik'}$,
(b) delete component k and set $z_i = k'$ with probability

$$\min\left\{1, \frac{n-1}{\alpha}\frac{h_i(k')p(y_i|x_i,\beta_{k'},\sigma^2_{k'})}{h_i(k)p(y_i|x_i,\beta_k,\sigma^2_k)}\right\}$$

The resulting number of non-empty components is the new value of K
Step 2 (update component memberships): for $i=1,\ldots,n$, if data point i belongs to a component with more than one occupant, update its component membership with probability equal to

$$\frac{n^\star_{-ik}h_i(k)p(y_i|x_i,\beta_k,\sigma^2_k)}{\sum_{k=1}^K n^*_{-ik}h_i(k)p(y_i|x_i,\beta_k,\sigma^2_k)};$$

otherwise do nothing
Step 3 (update components' mixture parameters): for $k=1,\ldots,K$, update the mixture parameters by sampling from the full conditionals
Step 4 (iteration): repeat steps 1–3 until convergence

mance compared with the presented model was negligible. Our experience in this respect agrees with that of Rasmussen (2000).

Before the application of the algorithm, an initial clustering needs to be made and all data points should be assigned to their components. We fix the maximum number of mixture components to $K_{\max}$ ($\leqslant n$). The first component is created by generating values for its parameters from their prior distributions. Next, the normal density value with these component parameters is calculated for all data points. The data point with the largest density value is assigned to the first component. The second component is generated in the same way as the first, but the second data point can be assigned to the first component or to a second component depending on the density values by using each component parameters. This continues until all data points have been assigned to a finite number K ($\leqslant K_{\max}$) of components.

In the initial clustering as well as after the MCMC sampler has swept through updates of component memberships and parameters as described in steps 1–4 (Table 1), there is a possibility of creating some components with very few data points. To avoid overfitting problems, one can optionally eliminate redundant components after convergence. In this study we eliminate components whose data points amount to less than 5% of the sample size. These data points are then transferred to existing components depending on their maximum normal density values.

Posterior modes are the natural choice for point estimation of the regression parameters. However, as the posterior modes might not be preserved under marginalization (Park and Casella, 2008), we use posterior medians to infer regression parameters of the components. The final number of components after the algorithm has converged, or after the optional elimination step if this step was added, is the estimate of the number of components K.

5. Simulation

In this section we assess the performance of the proposed method on various simulated data sets and demonstrate the method's ability in two major aspects. The first aspect concerns the data allocation strategy where we compare the proposed DICRP with the original CRP in the context of mixture regression models. Secondly, we assess the performance of DIMR to demonstrate its potential in estimation of the number of mixture components and the regression coefficients as well as in prediction of the response variable. Furthermore, we compare the performance of DIMR with similar estimation methods for mixture regression, namely FMR-Lasso (Städler *et al.*, 2010), the Bayesian mixture regression (BMR) method (Hurn *et al.*, 2003), glmnet (Friedman *et al.*, 2010) and RandomForest (Breiman, 2001). Finally, we examine the empirical consistency of DIMR for a fixed dimension. We used two data generation scenarios with different numbers of components and mixing probabilities: Table 2.

Throughout this section, the performance of our algorithm in terms of estimating regression coefficients is evaluated through mean-squared errors between the true parameter values and their corresponding estimates over all regression coefficients and mixture components. To investigate prediction errors we employ a fivefold cross-validation. For model comparison, we use the average of the mean-squared errors over the five test sets.

5.1. Clustering accuracy

We first performed a simulation study to asseses the clustering accuracy of the DICRP and to compare it with the original CRP. We also demonstrate the influence of clustering accuracy on fitting mixture regression models. For this, 100 independent data sets with $p = n = 100$ were generated from model 1 (Table 2). To generate the additional data $\mathbf{S}$, for each of the K clusters of simulated data we randomly selected $100\kappa\%$ of the data points in the cluster and set $s_{i,i'} = 1$ for each pair of points i and i' in the selected set. All other similarity values were set to 0. This means that the value of κ determined the level of informativeness of the generated $\mathbf{S}$. We considered five levels of informativeness $\kappa = 0, 0.1, 0.3, 0.6, 0.8$, where $\kappa = 0$ means that $S = 0$ which yields the original CRP.

After convergence of the MCMC algorithm we applied the optional elimination step described above and the final mixture clusters were obtained. This clustering was compared with the actual

Table 2. Simulation models

Model	$\boldsymbol{\pi}$	$\boldsymbol{\sigma}^2$	$\boldsymbol{\beta}$
1	(0.5,0.5)	(0.5,0.5)	$\beta_1 = (\mathbf{0}_{p-5}, 5, 5, 5, 5, 5)$ $\beta_2 = (\mathbf{0}_{p-5}, -3, -3, -3, -3, -3)$
2	(0.1,0.3,0.6)	(0.5,0.5,0.5)	$\beta_1 = (\mathbf{0}_{p-5}, 5, 5, 5, 5, 5)$ $\beta_2 = (\mathbf{0}_{p-5}, -3, -3, -3, -3, -3)$ $\beta_3 = (\mathbf{0}_{p-5}, 1, 1, 1, 1, 1)$

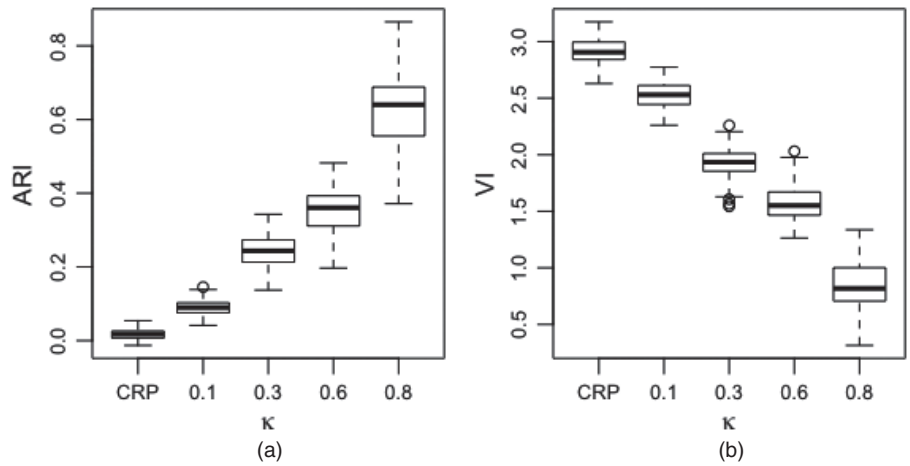

Fig. 2. Clustering accuracy of DICRP with varying informativeness parameter κ in terms of the (a) adjusted Rand index and (b) variation of information, under model 1 for 100 independent data sets with $p = n = 100$

clustering of the simulated data to determine the clustering accuracy. There are various measures of partition correspondence among which we chose two well-known measures, namely the Rand index ARI of Hubert and Arabie (1985) and the variation of information VI of Meilă (2005). ARI can take continuous values between 0 (for independent clustering) and 1 (for identical clustering), whereas VI is positive with 0 for identical clustering and grows when the distance between two clusterings becomes larger.

Fig. 2 presents the results of the clustering by using the original CRP and DICRP with varying levels of prior informativeness κ. Note that, as κ increases, both ARI (Fig. 2(a)) and VI (Fig. 2(b)) suggest continued gain in clustering accuracy. The first boxplot in both plots represents the clustering accuracy of the original CRP. As suggested by Fig. 2, the presence of more informative additional data can increase clustering accuracy significantly. Naturally, we expect higher accuracy in clustering to lead to more accurate mixture parameter estimation and better predictions. Fig. 3 indicates a clear increase in correctly estimating the number of mixture components (described as power in Fig. 3) (Fig. 3(a)) and decrease in both regression coefficient estimation errors (Fig. 3(b)) and prediction errors (Fig. 3(c)) when the informativeness of the additional data increases.

5.2. Comparison of methods

As mentioned above, we compared DIMR with several similar methods: FMRLasso, BMR, glmnet and randomForest. FMRLasso is a penalized maximum likelihood approach (Städler *et al.*, 2010), and BMR a similar Bayesian approach (Hurn *et al.*, 2003). Although they are alternatives for estimation of mixture regression models, they are not suitable for incorporation of additional data. Therefore, we also used glmnet and randomForest, which allow utilization of additional data, to compare with. Since the latter are of non-mixture regression model type, the comparison of these methods with DIMR is based only on model prediction errors. However, for comparison with FMRLasso and BMR we take into account various aspects such as the number of times that the number of components was correctly discovered (power), the number of false positive results, FP, the number of true positive results, TP, estimation errors of the regression coefficients and prediction errors.

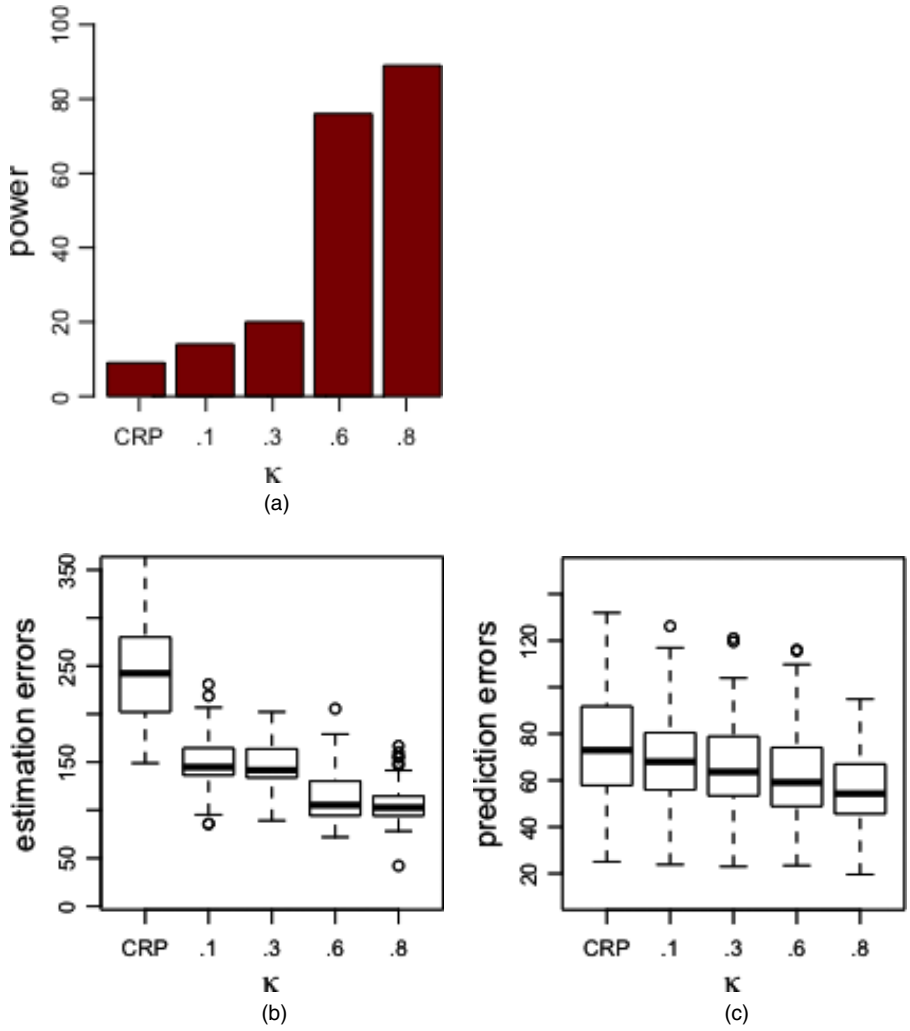

Fig. 3. Results of applying DIMR to simulated data under model 1 with various levels of informativeness κ for the additional data: evaluations are based on 100 simulation runs to calculate (a) the number of times that the number of mixture components was correctly estimated (power), (b) l_2-loss of all regression coefficients and mixture components and (c) the mean-squared error of response predictions

All five methods were applied to 100 independent data sets of size $n = 100$ that were generated from model 1 of Table 2. Comparing with FMRLasso and BMR we let the data dimension p range from 20 to 300, whereas for glmnet and randomForest we fixed the data dimension to $p = 20$ while letting the level of informativeness κ take values in $\{0, 0.3, 0.5, 0.8\}$. For the former case, calculation of estimation errors is possible only when the number of components equals the number of components of the simulated data set. Therefore, when either of the two methods misidentified the number of mixture components, new data sets were generated.

FMRLasso requires a predetermined number of components. For a given number of components, FMRLasso selects a penalty parameter out of a grid of proposal values based on BIC or AIC of the fitted models. We set the number of components to vary from 1 to 5 and fixed the grid of tuning parameters as suggested in Städler *et al.* (2010). For selection, we used BIC.

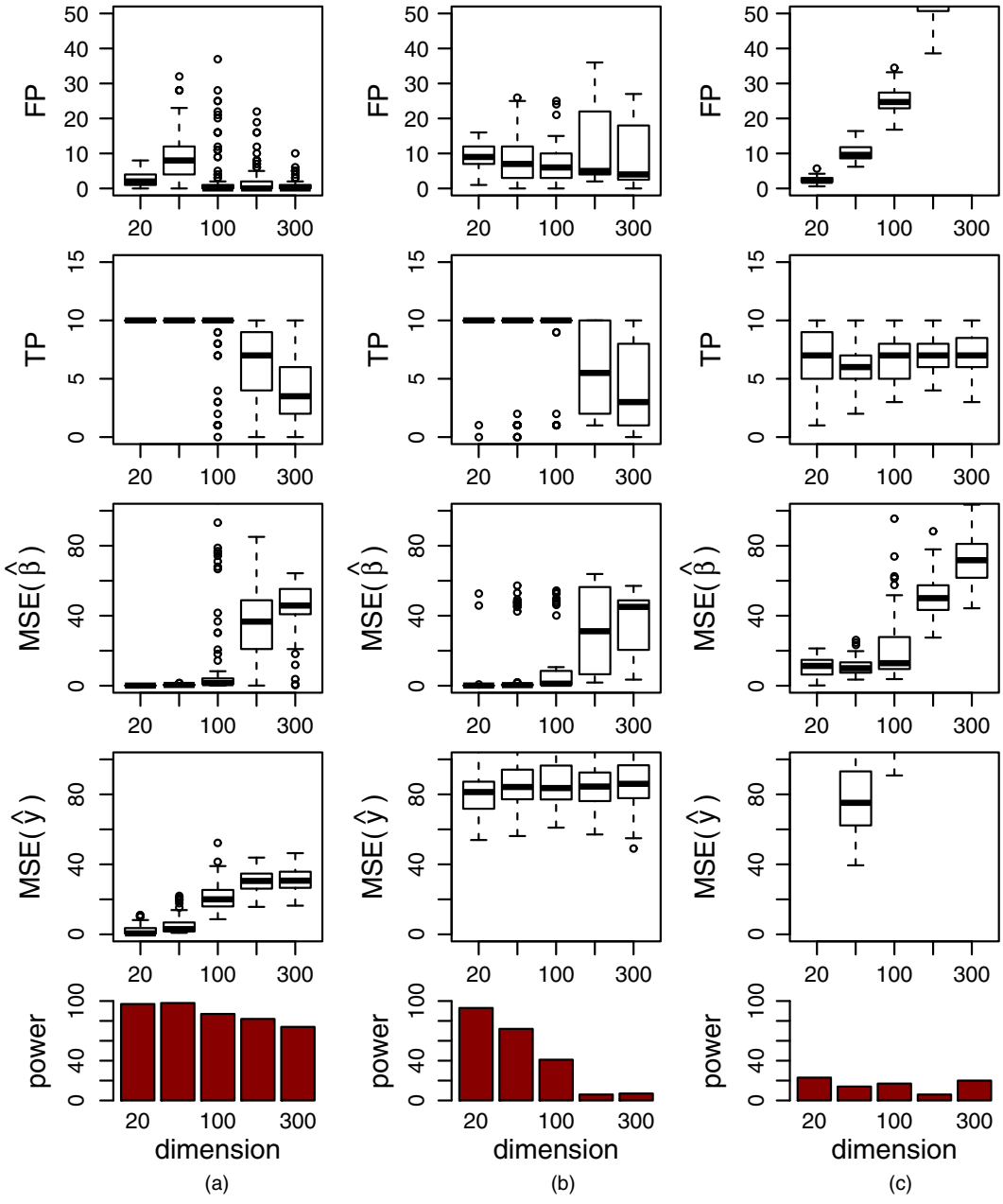

Fig. 4. Comparison of (a) DIMR, (b) FMRLasso and (c) BMR under model 1 in terms of false positive and true positive results, estimation and prediction errors, and correctly detected number of components (i.e. power)

For DIMR the additional data were plugged in as described in Section 5.1. For incorporation of the additional data in glmnet and randomForest we cannot work with similarity data. Instead, we partitioned the simulated clusters of main data based on the value of the predefined level of informativeness κ and fitted the model through separate linear regression models using the two methods. More precisely, for each of the $K = 2$ clusters of simulated data we randomly

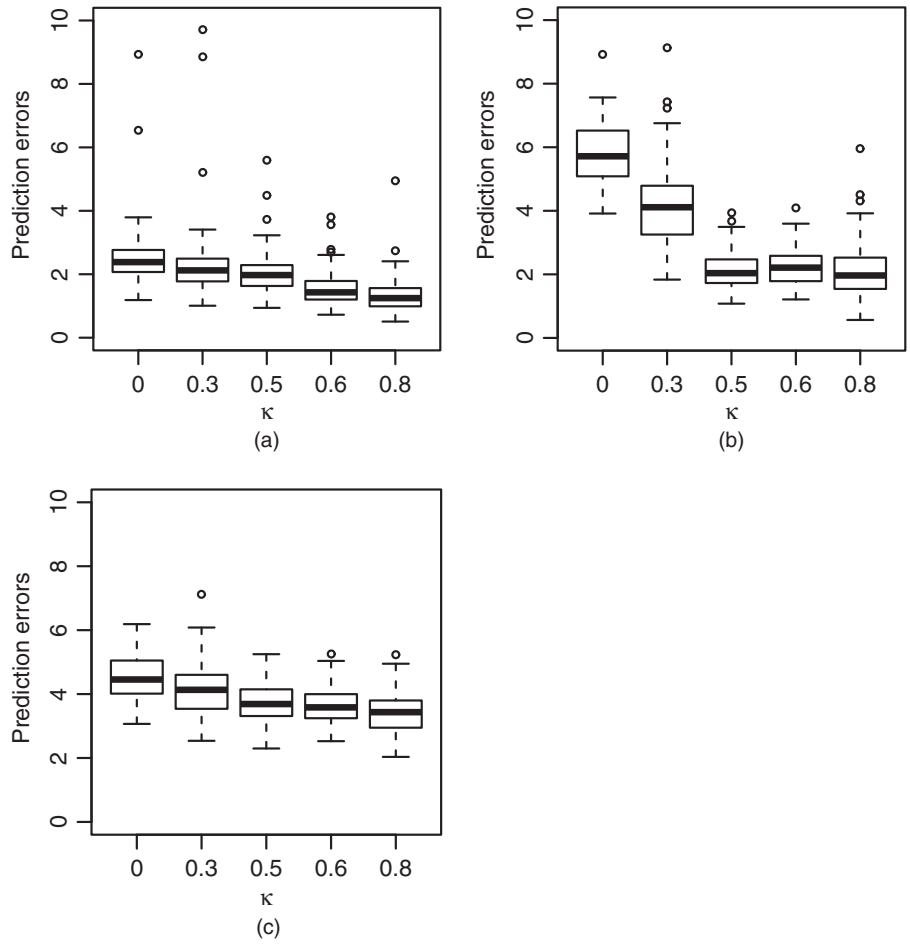

Fig. 5. Comparison of (a) DIMR, (b) glmnet and (c) randomForest under model 1 in terms of prediction errors for varying level of informativeness of the additional information

selected $100\kappa\%$ of the data points in the cluster. We then performed regression analysis and calculated cross-validated prediction errors based on three data splits: the two selected sets and the set consisting of the remaining data points. For $\kappa = 0$ there was no split and regression was performed on all data points.

Fig. 4 shows comparable performance for the three mixture methods DIMR, FMRLasso and BMR in terms of regression coefficient estimation. In higher dimensions, however, we see a better performance for DIMR and FMRLasso compared with BMR. Similarly, for model selection, especially in higher dimensions, the performances of DIMR and FMRLasso are much more reasonable than that of BMR. This is somewhat expected as model selection is not properly addressed in BMR.

Concerning model prediction, there is a significant gap between the performance of the three mixture methods. This can be explained by the fact that DIMR shows a more successful performance in terms of detecting the true number of mixture components as can be seen in the three bottom boxplots in Fig. 4. For BMR, this might be mainly due to overfitting caused by the way that the number of components is selected, which is through comparing the posterior

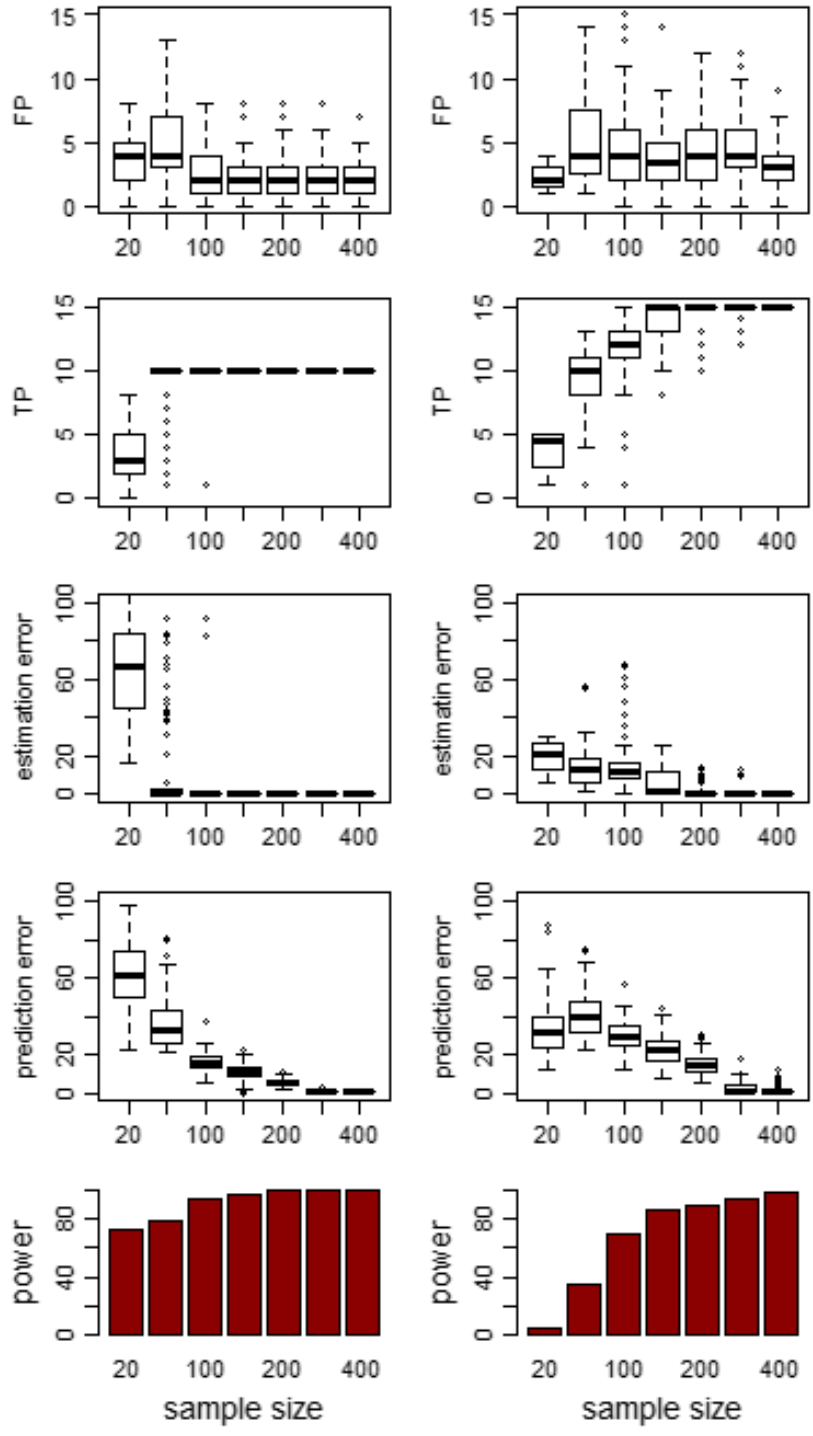

Fig. 6. Consistency of DIMR under (a) model 1 and (b) model 2 in terms of false positive and true positive results, prediction and estimation errors and power

probability of the varying k. In contrast, FMRLasso employs a BIC-approach to select the number of mixture components. Therefore, a higher data dimension obviously leads to heavier penalization of the log-likelihood and favouring a smaller number of components.

Fig. 5 compares the performance of the three data integrative methods DIMR, glmnet and randomForest based on varying levels of informativeness utilized for model estimations. As expected, the performance of all three methods improves with an increased level of informativeness of the additional data. Generally, the performance of DIMR is better than that of the two other methods.

5.3. Empirical consistency

To demonstrate the consistency of our estimator, we considered six different sample sizes varying from $n = 20$ to $n = 400$. For each sample size 100 independent data sets are generated with a fixed dimension $p = 100$. Here we are interested in comparing empirical consistency results when the complexity of the data is enhanced by an increase in the number of mixture components with unbalanced mixture probabilities. Therefore, in addition to simulations from model 1, we generated data from model 2 in Table 2.

Fig. 6 illustrates how a larger sample size can improve the performance of DIMR. Particularly, estimation and prediction errors tend towards 0 when the sample size increases. This holds for the simulated data sets under both models, but the convergence is slower for the more complex data from model 2.

6. Application

We consider an application of DIMR and two other regression methods, glmnet and randomForest, on a real data set of stomach adenocarcinoma (Cancer Genome Atlas Research Network, 2014). This data set contains various measurements on the deoxyribonucleic acid of 260 patients among which we are interested in messenger ribonucleic acid (RNA) measurements of genes. More specifically, we shall focus on messenger RNA measurements of one specific gene, namely CDH1, that is suggested to be associated with gastric cancer (Keller *et al.*, 1996). Among 1357 genes, we selected 298 genes that present high correlations with CDH1 as predictor variables. Our major aim is to project messenger RNA measurement of CDH1 on the other genes, yet trying to capture heterogeneity among the samples.

Stomach adenocarcinoma is the most frequent type of gastric cancer that is classified into different subtypes. The most used classification is the so-called Lauren classification with two classes: diffuse and intestinal type. Another popular classification concerns the World Health Organization classification with four clusters, namely papillary, tubular, mucinous (colloid) and poorly cohesive carcinomas (Bosman *et al.*, 2010). Besides these classification systems, in a recent study of gastric adenocarcinoma a molecular classification is proposed that suggests four subtypes that have more clinical utilities (Cancer Genome Atlas Research Network, 2014). This raises the question whether or not these external classification data can be a useful asset in capturing heterogeneity of the genomic data in terms of reduction in the prediction errors.

We used the data from the three above-mentioned classification schemes together with the micro-RNA expression clusters and methylation clusters as the sources of additional information. We employed each regression method with and without additional data. For DIMR, these additional data were incorporated in the models either individually or jointly in the form of a similarity matrix $\mathbf{S}$. To form a similarity matrix based on an individual additional source of data we set $s_{ii'} = 1$ when observations i and i' are clustered together, and $s_{ii'} = 0$ otherwise. To form a similarity matrix that is based on more than one source of data, we used an *overlap* measure

Table 3. Additional data incorporation scenarios for gastric adenocarcinoma†

Scenario	*Lauren classification*	*Micro-RNA expression*	*Methylation clusters*	*Molecular classification*	*World Health Organization classification*	*off-Diag S*
I	×	×	×	×	×	$s_{ii'}=0$
II	✓	×	×	×	×	$s_{ii'}=1$
III	×	✓	×	×	×	$s_{ii'}=1$
IV	×	×	✓	×	×	$s_{ii'}=1$
V	×	×	×	✓	×	$s_{ii'}=1$
VI	×	×	×	×	✓	$s_{ii'}=1$
VII	✓	✓	×	✓	×	$s_{ii'}=3$
VIII	✓	✓	✓	✓	✓	$s_{ii'}=5$

†We generate eight similarity matrices S each containing information from none (first row), one (second–sixth rows) or multiple additional data sources (seventh and eighth rows). We assume an overlap measure for which $s_{ii'}=0$ if i and i' do not belong to the same cluster in the corresponding additional data which are marked by '×'. Otherwise, based on the number of data sources in which i and i' belong to the same cluster (marked by '✓'), $s_{ii'}$ can take values 1, 3 or 5.

which means that $s_{ii'}=k$ when in k additional sources of data the two observations are in the same cluster. Here we consider eight scenarios, the first of which assumes no additional data (i.e. $\mathbf{S}=0$). In this case, as mentioned earlier, the data allocation scheme DICRP is equivalent to the original CRP. The next five scenarios concern generation of five similarity matrices each using one additional source of data separately. The last two similarity matrices use integration of three and five additional sources of data. Table 3 gives a summary of the various data integration scenarios and generation of their corresponding similarity matrices.

A different approach has been taken to incorporate the additional data in the glmnet and randomForest methods. We partitioned the main data by using clusters based on the additional data and fitted separate linear regression models by using these methods. To be fair in the comparison, new clusterings were built from combining the additional data and using a k-mode algorithm (Huang, 1997).

We evaluated the performance of the methods in terms of model prediction by using a fivefold cross-validation method. Fig. 7 presents the effectiveness of the additional data sets when they are used as the only external source and of some combinations. Clearly, when the additional data set that is used in the algorithm comprises information that is obtained from a combination of additional data sets, the prediction errors dramatically decrease. This is even more obvious from the first panel in Fig. 8 which particularly shows unreliable predictions when clusters of micro-RNA expression are used as the only external source, and a considerable mitigation of the prediction errors when these data are added as additional data in the DIMR algorithm. This improvement in model estimation might be due to the complementary role of the additional sources of data and the suitability of the DIMR algorithm for the integration of such data. Fig. 8 also shows a comparable performance for the glmnet and the randomForest methods. Yet, as can be seen from the boxplots, DIMR very often outperforms the other two methods, especially when different sources of additional data are used.

7. Discussion

In this paper we developed a Bayesian approach accompanied by an MCMC algorithm, DIMR, for mixture regression estimation which provides a flexible framework to incorporate auxiliary

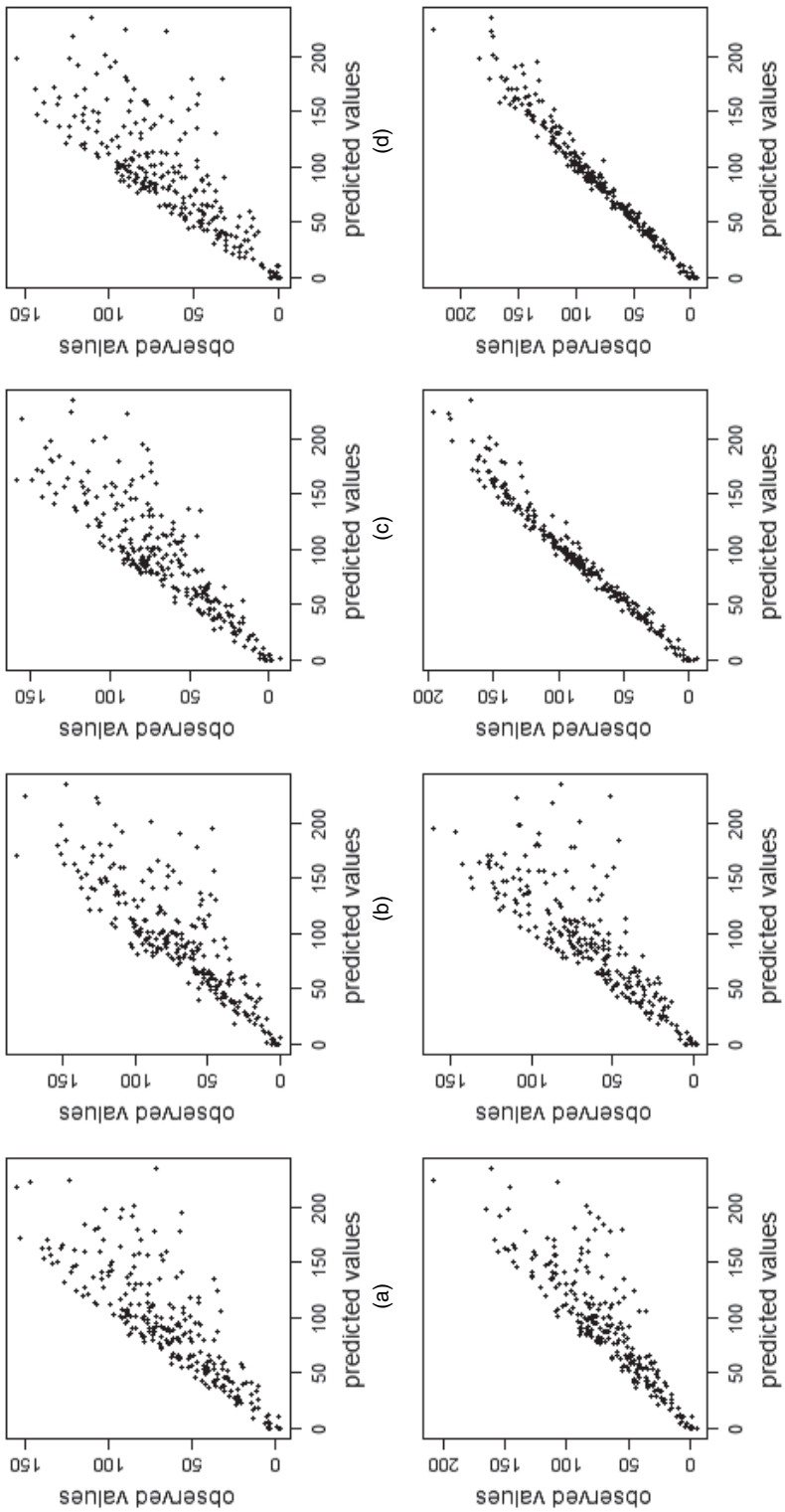

Fig. 7. Prediction *versus* response plots of fivefold cross-validated DIMR by using various additional sources: (a) no external data; (b) Lauren classification data; (c) micro-RNA expression data; (d) methylation clusters data; (e) molecular classification data; (f) World Health Organization classification data; (g) Lauren classification and molecular classification and micro-RNA expression data; (h) all additional data

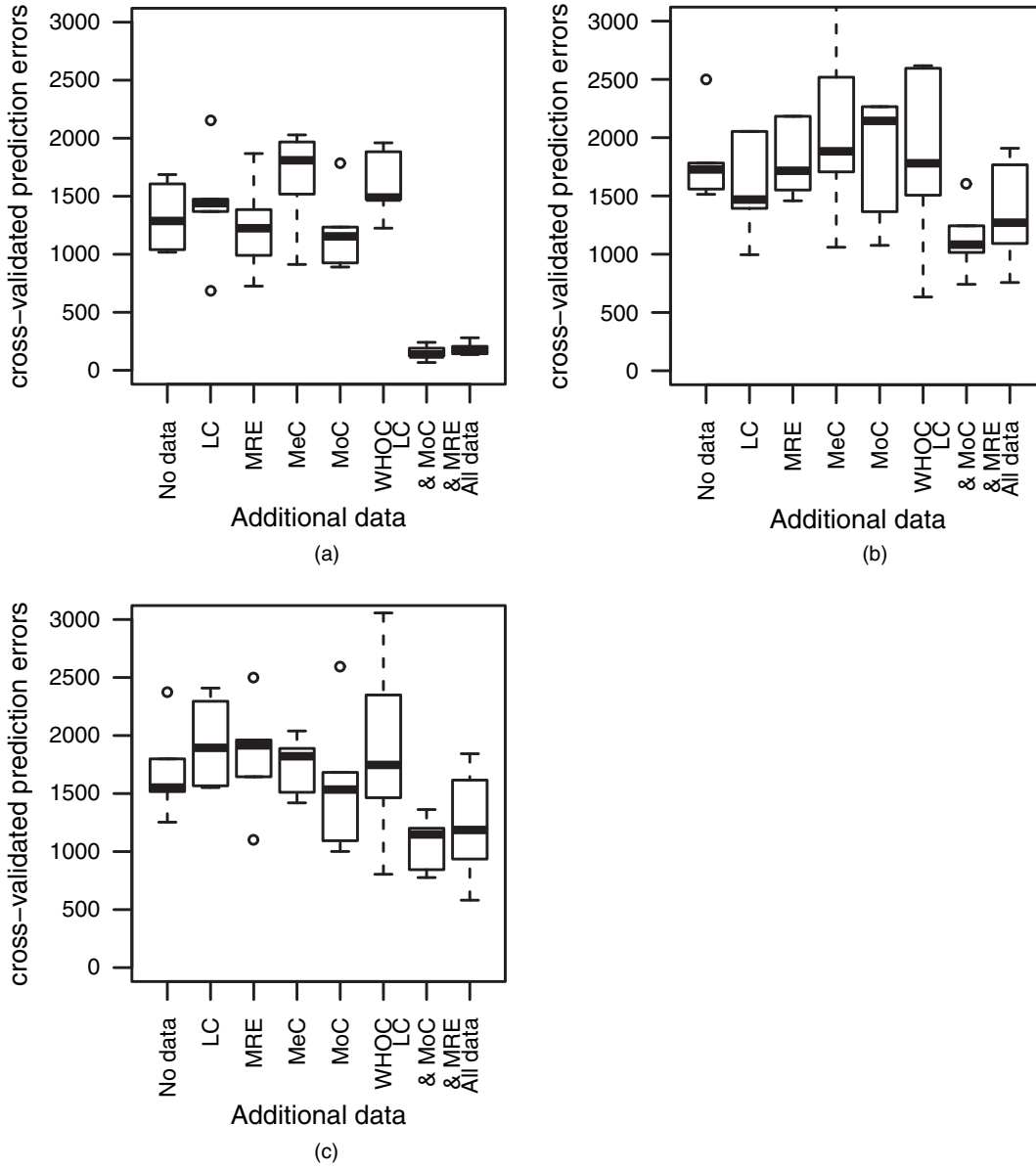

Fig. 8. Cross-validated predicted mean-squared error of regression models based on various additional sources of data for stomach cancer ('no data' refers to the situation where the methods are given no additional data, and 'all data' to the case where integration of all additional data sets is taken into account): (a) DIMR; (b) glmnet; (c) randomForest

information. The method aims at facilitating the analysis of reasonably high dimensional data sets through considering shrinkage-type priors on the regression coefficients. We further extended the CRP to a more efficient data allocation scheme DICRP, that is placed in the heart of our algorithm, yet could be of interest in other problems independently. The performance of the method was investigated through an extensive simulation study and by application of the method to real data. The results demonstrate that DICRP is more successful than the CRP

when additional data on the similarity of the data points are available. Furthermore, the results from the comparative study demonstrate that our approach is competitive with the penalized likelihood method FMRLasso, and generally outperforms BMR.

In this work we combined an MCMC algorithm with a Gibbs sampler to infer mixture and regression parameters jointly. Although MCMC methods have good properties, for large data sets computational costs can be high. A useful alternative for MCMC sampling is variational inference, which is based on maximization of the marginal likelihood (Blei and Jordan, 2006; Ma and Leijon, 2011; Fan *et al.*, 2012; Ma *et al.*, 2014). Variational inference methods are generally faster than MCMC methods and it may be worthwhile to investigate them in conjunction with our proposed data integration scheme.

The Bayesian lasso does not yield sparse solutions directly. This is why in this paper sparsity is induced from model selection through credible intervals of the posterior samples. As shown in Section 4, utilizing this method within our mixture framework yields better selection results compared with similar methods. However, considering alternative selection methods such as Zhao and Sarkar (2015) or Carvalho *et al.* (2010) might be useful for further improvement of the selection procedure.

Mixture models are identifiable only up to a permutation of the component labels. For sampling approaches this only affects interpretation of results but is no problem for parameter estimation itself (Celeux *et al.*, 2000; Jasra *et al.*, 2005). We endeavour to tackle this problem by adapting a unique labelling for the components based on ordering the l_1-norm of the regression coefficients $|\beta_1| < |\beta_2| < \ldots < |\beta_K|$. Although this assumption in the MCMC algorithm avoids numerical (label switching) problems, a general identifiability problem, i.e. the correct identification of the components that better describe the response variable, still remains an open question to investigate (Papastamoulis and Iliopoulos, 2010; Rodríguez and Walker, 2014).

In practice, the similarity matrix either is given or must be constructed on the basis of independently available information which could be of continuous or categorical nature, or both. In case the similarity matrix must be constructed, different approaches could be used to translate additional data into a similarity matrix. Which approach is chosen is highly dependent on the domain and the application and to a certain extent subjective. For our simulation study it was sufficient to consider the similarity matrix as given. For our real data example, the additional sources of data were all variables of the categorical type and the similarity matrix had to be computed from these categorical variables; the overlap measure that we used for this is a natural measure. However, depending on the type of additional information other similarity functions might be investigated (see, for example, measures provided in Boriah *et al.* (2008) and Cha (2007)). Similarly, instead of an overlap function that summarizes the similarity values stemming from different sources of data, other choices could be explored. Investigation of different similarity measures is beyond the scope of the present paper. In a forthcoming paper we study various types of similarity measures and summarization methods based on both discrete and continuous types of additional attributes.

Acknowledgements

The authors are grateful to the Associate Editor and a referee for their helpful suggestions.

The authors have declared no conflict of interest.

References

Antoniak, C. E. (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.*, **1**, 1152–1174.

Blei, D. M. and Frazier, P. (2011) Distance dependant Chinese restaurant process. *J. Mach. Learn. Res.*, **12**, 2461–2488.
Blei, D. M. and Jordan, M. I. (2006) Variational inference for Dirichlet process mixtures. *Baysn Anal.*, **1**, 121–143.
Boriah, S., Chandola, V. and Kumar, V. (2008) Similarity measures for categorical data: a comparative evaluation. In *Proc. Int. Conf. Data Mining*, pp. 243–254. Philadelphia: Society for Industrial and Applied Mathematics.
Bosman, F. T., Carneiro, F., Hruban, R. H. and Theise, N. D. (2010) *WHO Classification of Tumours of the Digestive System*, no. 4. Geneva: World Health Organization.
Breiman, L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.
Cancer Genome Atlas Research Network (2014) Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, **513**, 202–209.
Carvalho, C. M., Polson, N. G. and Scott, J. G. (2010) The horseshoe estimator for sparse signals. *Biometrika*, **97**, 465–480.
Celeux, G., Hurn, M. and Robert, C. P. (2000) Computational and inferential difficulties with mixture posterior distributions. *J. Am. Statist. Ass.*, **95**, 957–970.
Cha, S.-H. (2007) Comprehensive survey on distance/similarity measures between probability density functions. *City*, **1**, 300–307.
Chung, Y. and Dunson, D. B. (2009) Nonparametric Bayes conditional distribution modeling with variable selection. *J. Am. Statist. Ass.*, **104**, 1646–1660.
Escobar, M. and West, M. (1995) Bayesian prediction and density estimation. *J. Am. Statist. Ass.*, **90** 577–588.
Everitt, B. S. and Hand, D. J. (1981) *Finite Mixture Distributions*. London: Chapman and Hall.
Fan, W., Bouguila, N. and Ziou, D. (2012) Variational learning for finite Dirichlet mixture models and applications. *IEEE Trans. Neurl Netwrks Learn. Syst.*, **23**, 762–774.
Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Statist. Softwr.*, **33**, 1–22.
Goldfeld, S. M. and Quandt, R. E. (1976) A Markov model for switching regression. *J. Econmetr.*, **1**, 3–16.
Gupta, M. and Ibrahim, J. G. (2007) Variable selection in regression mixture modeling for the discovery of gene regulatory networks. *J. Am. Statist. Ass.*, **102**, 867–880.
Huang, Z. (1997) A fast clustering algorithm to cluster very large categorical data sets in data mining. In *Data Mining and Knowledge Discovery*, vol. 3, pp. 34–39.
Hubert, L. and Arabie, P. (1985) Comparing partitions. *J. Classificn*, **2**, 193–218.
Hurn, M., Justel, A. and Robert, C. P. (2003) Estimating mixtures of regressions. *J. Computnl Graph. Statist.*, **12**, 55–79.
Jacobs, R. A., Jordan, M. I., Nowlan, S. J. and Hinton, G. E. (1991) Adaptive mixture of local experts. *Neurl Computn*, **3**, 79–87.
Jacobs, R. A., Peng, F. and Tanner, M. A. (1997) A Bayesian approach to model selection in hierarchical mixtures-of-experts architectures. *Neurl Netwrks*, **10**, 231–241.
Jasra, A., Holmes, C. C. and Stephens, D. A. (2005) Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statist. Sci.*, **20**, 50–67.
Jordan, M. I. and Jacobs, R. A. (1994) Hierarchical mixtures of experts and the EM algorithm. *Neurl Computn*, **6**, 181–214.
Keller, G., Grimm, V., Vogelsang, H., Bischoff, P., Mueller, J., Siewert, J. R. and Höfler, H. (1996) Analysis for microsatellite instability and mutations of the DNA mismatch repair gene HMLH1 in familial gastric cancer. *Int. J. Cancer*, **68**, 571–576.
Khalili, A. (2011) An overview of the new feature selection methods in finite mixture of regression models. *J. Iran. Statist. Soc.*, **10**, 201–235.
Khalili, A. and Chen, J. (2007) Variable selection in finite mixture of regression models. *J. Am. Statist. Ass.*, **102**, 1025–1038.
Ma, Z. and Leijon, A. (2011) Bayesian estimation of beta mixture models with variational inference. *IEEE Trans. Pattn Anal. Mach. Intell.*, 2160–2173.
Ma, Z., Rana, P. K. Taghia, J., Flierl, M. and Leijon, A. (2014) Bayesian estimation of Dirichlet mixture model with variational inference. *Pattn Recogn*, **47**, 3143–3157.
Maugis, C., Celeux, G. and Martin-Magniette, M.-L. (2009) Variable selection for clustering with Gaussian mixture models. *Biometrics*, **65**, 701–709.
McLachlan, G. and Peel, D. (2000) *Finite Mixture Distributions*. New York: Wiley.
Meilă, M. (2005) Comparing clusterings: an axiomatic view. In *Proc. 22nd Int. Conf. Machine Learning*, pp. 577–584. New York: Association for Computing Machinery Press.
Müller, P. and Quintana, F. (2010) Random partition models with regression on covariates. *J. Statist. Planng Inf.*, **140**, 2801–2808.
Neal, R. M. (2000) Markov chain sampling methods for Dirichlet process mixture models. *J. Computnl Graph. Statist.*, **9**, 249–265.
Papastamoulis, P. and Iliopoulos, G. (2010) An artificial allocations based solution to the label switching problem in Bayesian analysis of mixtures of distributions. *J. Computnl Graph. Statist.*, **19**, 313–331.
Park, T. and Casella, G. (2008) The Bayesian lasso. *J. Am. Statist. Ass.*, **103**, 681–686.

Quandt, R. and Ramsey, J. (1978) Estimating mixtures of normal distributions and switching regression. *J. Am. Statist. Ass.*, **73**, 730–738.
Rasmussen, C. E. (2000) The infinite Gaussian mixture model. In *Neural Information Processing Systems*, vol. 12, pp. 554–560.
Rasmussen, C. E. and Ghahramani, Z. (2002) Infinite mixtures of Gaussian process experts. *Adv. Neurl Inform Process. Syst.*, **2**, 881–888.
Richardson, S. and Green, P. J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. R. Statist. Soc.* B, **59**, 731–792; correction, **60** (1998), 661.
Rodríguez, C. E. and Walker, S. G. (2014) Label switching in Bayesian mixture models: deterministic relabeling strategies. *J. Computnl Graph. Statist.*, **23**, 25–45.
Städler, N., Bühlmann, P. and Van De Geer, S. (2010) ℓ_1-penalization for mixture regression models. *Test*, **19**, 209–256.
Stephens, M. (2000) Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods. *Ann. Statist.*, **28**, 40–74.
Tadesse, M. G., Sha, N. and Vannucci, M. (2005) Bayesian variable selection in clustering high-dimensional data. *J. Am. Statist. Ass.*, **100**, 602–617.
Tibshirani, R. (2011) Regression shrinkage and selection via the lasso: a retrospective. *J. R. Statist. Soc.* B, **73**, 273–282.
Tran, M.-N., Nott, D. J. and Kohn, R. (2012) Simultaneous variable selection and component selection for regression density estimation with mixtures of heteroscedastic experts. *Electron. J. Statist.*, **6** 1170–1199.
Villani, M., Kohn, R. and Giordani, P (2009) Regression density estimation using smooth adaptive Gaussian mixtures. *J. Econmetr.*, **153**, 155–173.
Wedel, M. and Kamakura, W. A. (2012) *Market Segmentation: Conceptual and Methodological Foundations*, vol. 8. New York: Springer Science and Business Media.
Williams, P. M. (1995) Bayesian regularization and pruning using a Laplace prior. *Neurl Computn*, **7**, 117–143.
Xie, J., Ma, Z., Zhang, G., Xue, J.-H., Chien, J.-T., Lin, Z. and Guo, J. (2018) Balson: Bayesian least squares optimization with nonnegative l1-norm constraint. In *Proc. 28th Int. Wrkshp Machine Learning for Signal Processing*, pp. 1–6. New York: Institute of Electrical and Electronics Engineers.
Yau, C. and Holmes, C. (2011) Hierarchical Bayesian nonparametric mixture models for clustering with variable relevance determination. *Baysn Anal.*, **6**, 329–352.
Zhao, Z. and Sarkar, S. K. (2015) A Bayesian approach to constructing multiple confidence intervals of selected parameters with sparse signals. *Statist. Sin.*, **25**, 725–741.
Zou, H. (2006) The adaptive lasso and its oracle properties. *J. Am. Statist. Ass.*, **101**, 1418–1429.